NeurIPS 2019: Pseudo-Extended Markov chain Monte Carlo (paper ID: 2415)

We would like to thank the reviewers for dedicating their time to review our paper and the helpful feedback they have provided to improve the quality of this work. All of the reviewers' minor comments and corrections have been added to the paper. Below, we address the reviewers' main questions.

Reviewer 1

1. ***The paper focuses on HMC sampling. Do you believe that PE can be used in MH sampling or in SMC sampling? With discrete variables?*** The pseudo-extended approach modifies the target and not the sampler and so other sampling algorithms (e.g. MH and SMC) could be used to replace the original target distribution with the pseudo-extended target. As for discrete variables, sampling from discrete state-spaces can be challenging due to issues of dimensionality (e.g. Boltzmann example in 4.2). There is a similar motivation for concrete distributions. Unfortunately, HMC can't be applied in the discrete setting due to discontinuous gradients. However, transforming the problem into a continuous space (e.g. Boltzmann relaxation) means that we can use efficient gradient-based MCMC algorithms like HMC to explore the posterior space efficiently.
2. ***Could you provide intuition on what role $\pi(\beta)$ and $g(\beta)$ play, with respect to characteristics of the inference problem? From Fig. 2, β plays a key role. How do you recommend setting π and g to best estimate β?*** This is a good question and something we have now expanded on in the paper. $\pi(\beta)$ is a prior on the tempering and in the experiments we assumed a uniform prior, however, in practice the user may have prior information about the multi-modality of the target, for example, whether the modes are far apart or tightly packed together, and so could choose a Beta prior which places more prior mass closer to 1 or 0, respectively. As for g, there's a lot of scope in future research to explore how this can be used to improve the mixing of the β parameters. For example, something that we have recently experimented with is $g(\beta) = \mathcal{N}(0, \Sigma)$, where $\Sigma_{ij} > -1/(N-1)$ and $\Sigma_{ii} = 1$. This gives a positive definite matrix for Σ where the β pseudo-samples are negatively correlated and so repel each other allowing for better exploration of the β-space.
3. ***Could you please explain briefly how the models you wrote in Stan ensure that the sampler implements Algorithm 1 from Appendix B?*** Essentially, the Stan software applies HMC sampling to a specified log-target density function. Therefore, it's quite straightforward to implement pseudo-extended HMC within Stan by replacing the original log-target density with the pseudo-extended log-target density. Note that we could use Algorithm 1 directly for standard HMC (i.e. without Stan) and tune the step-size and length-scale parameters manually.
4. ***Looking through the Stan programs, the PE sampling code tends to be longer and more complicated. Have you thought about ways to reduce the complexity of the sampling code?*** From the Stan files you may notice that the pseudo-extended implementation is quite similar across each of the models. We should be able to simplify the code significantly by creating a "pseudo-extended" function at the top of the Stan file and call this function in place of the usual log-target, i.e. `target+=pseudo-extended()`. We will implement this in the version of the code that we release on Github.

Reviewer 2

1. ***As a minor comment in line 58, it would be good to state that delta is an arbitrary differentiable function.*** This is a good point and we've corrected this in the paper.

Reviewer 3

1. ***The experiments in 4.1 and 4.2 use the RMSE error of the target variables which is quite unusual. The more appropriate accuracy measure is the predictive likelihood of held out data vs inference time. Further, there is no conditioning on data in 4.1 and 4.2 so these are very trivial experiments in that all we are trying to do is generate a sample from the prior.*** The advantage of using the RMSE in 4.1 and 4.2 is that for these examples we can calculate the expectations exactly which provides us with ground-truth for comparison. We were also aiming to reflect what authors of previous works have also reported. We agree that predictive accuracy is important and we report this in 4.3. By considering both challenging simulated examples (4.1 and 4.2) and real-data scenarios (4.3) we can explore a wider range of posterior behaviour while also linking directly with previously considered models from the literature.
2. ***Report predictive log likelihood of held out data versus time. (For example see the Stan ADVI paper which has various graphs of this style).*** We've looked at the Stan ADVI paper and plots of this type will be replicated in our paper.
3. ***The paper should give some suggestions to tune the hyperparameter N.*** This is a good point and we've added some discussion on selecting N in the manuscript.